
Challenges and Pitfalls in the Evaluation of Structural Variants Callers

Luca Denti*, Thomas Krannich, Tomáš Vinař, Rayan Chikhi,
Paola Bonizzoni, Broňa Brejová, Fereydoun Hormozdiari

**Comenius University in Bratislava*

⚡ RECOMBSEQ2026 - Poster lightning talk ⚡



UNIVERZITA
KOMENSKÉHO
V BRATISLAVE



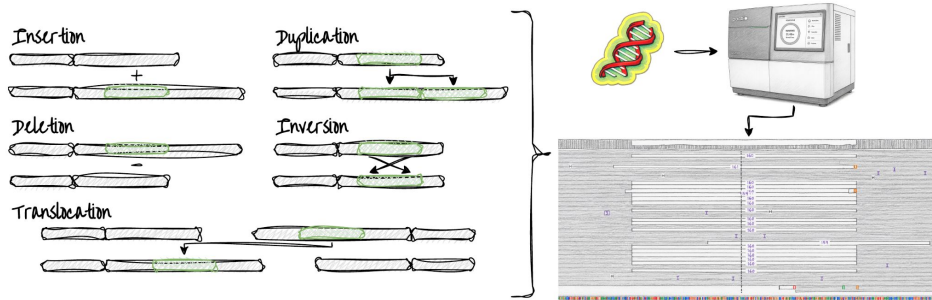
GERMAN
CANCER RESEARCH CENTER
IN THE HELMHOLTZ ASSOCIATION



UC DAVIS
UNIVERSITY OF CALIFORNIA

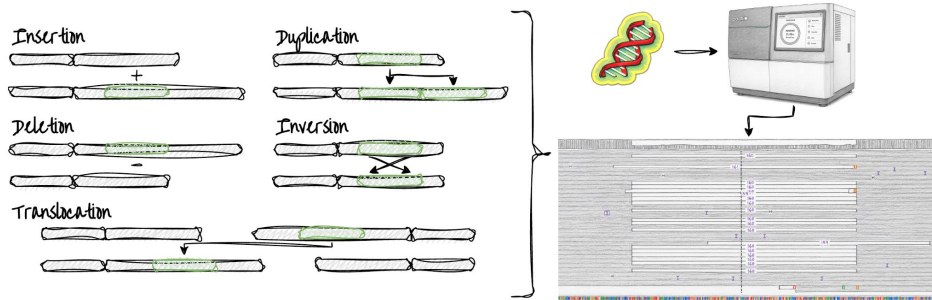
SV callers evaluation

1. Select reference genome
2. Call SVs from sequencing data
3. Select “ground truth” (*curated or created*)
4. Compare calls
 - a. *breakpoint-based + (optional) sequence realignment*
 - b. *sequence-based (uncommon)*



SV callers evaluation

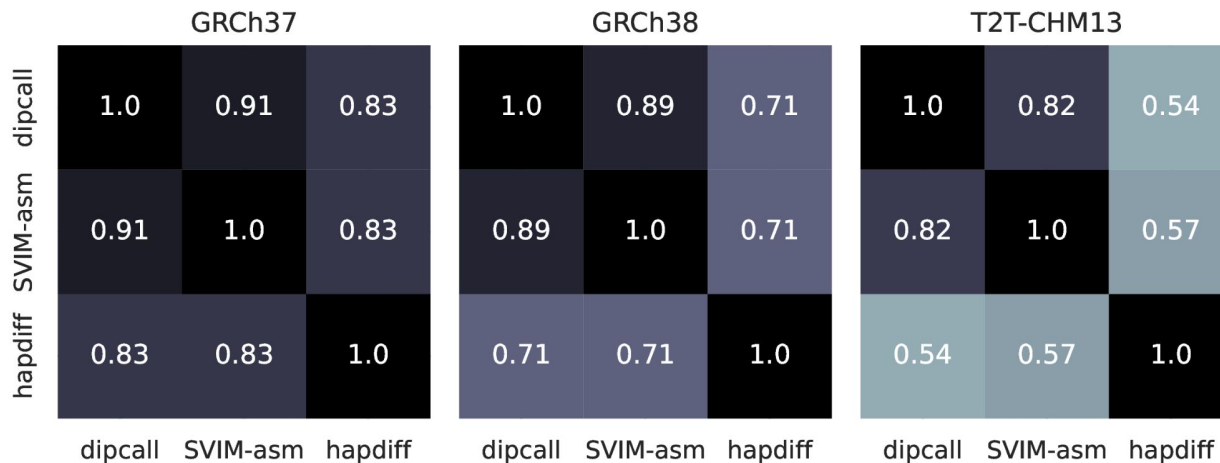
1. Select reference genome
2. Call SVs from sequencing data
3. Select “ground truth” (*curated or created*)
4. Compare calls
 - a. *breakpoint-based + (optional) sequence realignment*
 - b. *sequence-based (uncommon)*



Inconsistent benchmarking across studies ⇒ **contradictory findings**

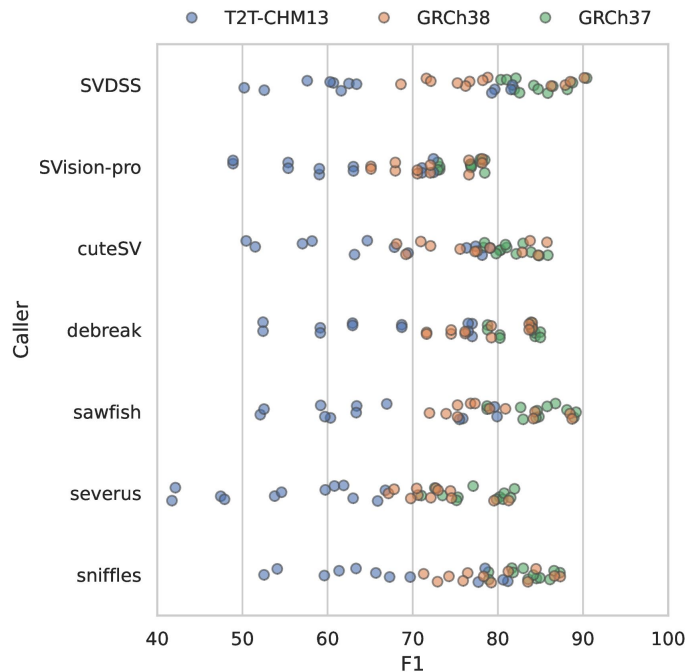
(our results reveal **expected but overlooked pitfalls** in current evaluation practices)

Assembly-based callers



Leading methods yielded different callsets, even when applied to the same assembly

Read-based callers



Performance depends on ground truth, reference, and benchmark parameters, even when applied to same sample

Poster

- User-defined choices cause overlooked divergence.
- Divergence is more evident in complex genomic regions.
- Results are not readily transferable across ground truths.
- Evaluation against a single ground truth is insufficient.
- No caller was best in every tested scenario.

Challenges and Pitfalls in the Evaluation of Structural Variants Callers

Luca Denti¹, Thomas Krannich², Tomáš Vinař³, Rayan Chikhi³, Paola Bonizzoni⁴, Broňa Brejová⁵, Fereydoun Hormozdizari⁵

¹ Comenius University in Bratislava, Slovakia; ² German Cancer Research Center, Germany; ³ Institut Pasteur, France

⁴ University of Milano-Bicocca, Italy; ⁵ University of California-Davis, USA



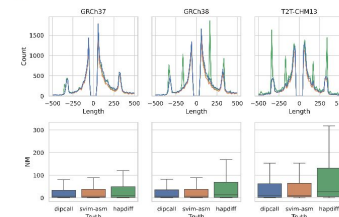
Motivation

Structural variants (SVs) are medium- and large-scale genomic alterations that shape phenotypic diversity and disease risk. Numerous methods have been proposed for the discovery of SVs, but their benchmarking has been inconsistent across studies, often resulting in *conflicting findings*. Primary sources of conflicting results lie in the inconsistent selection of reference genome version and ground truths, e.g., callers trained by consensus and all loci callers created from high-quality assemblies. Best practices and optimal evaluation strategies remain undefined and benchmarking results often vary, leaving the current performance assessment *difficult to interpret*. More importantly, this has created *ambiguity* about the true capabilities of testing methods to accurately detect SVs, with far-reaching implications for routine genome analysis.

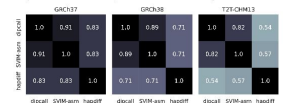
SV calling from high-quality de novo assemblies

HG002 (GIAB Q100), **HG005 (HPRC)**, **NA12878 (Platinum Pedigree)**

(I) Loading methods produced different callers, even when applied to the same assembly.

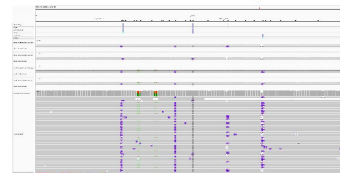


(II) Similarity decreases as the completeness of the reference genome increases.

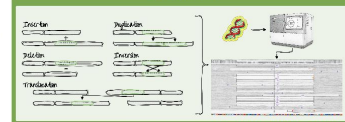


(III) Differing computational strategies applied to address the biological complexity of certain genomic loci are one of the primary sources of these discrepancies:

- alignment choices
- signal clustering (within and across haplotypes)



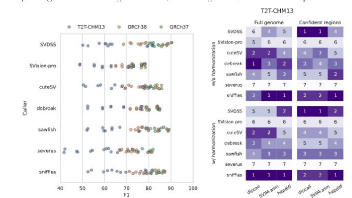
Background



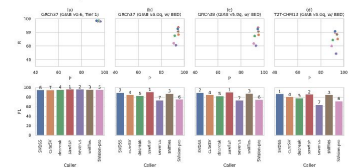
SV calling from PacBio HiFi reads

HG002 (Bio, HPRC), **NA12878 (Bio, Platinum Pedigree)**

(I) When evaluated against assembly-based callers, F1-scores and rankings of callers vary depending on the choice of ground truth, reference genome, and benchmark parameters.



(II) Similar results can be observed when considering curated SV callers. Remarkably, the inclusion of more (challenging) SVs located in complex genomic regions substantially affects SV callers' performance.



Discussion

- Common user-defined choices cause overlooked and unexpected divergence in performance evaluations.
- Performance results are not readily transferable between ground truths without careful consideration of their subtle differences.
- All long-read SV callers struggle more in complex genomic regions. Both calling and benchmarking are more challenging (due to inconsistent SV representation).
- Although some callers perform well across our benchmarks, no tool was best in every tested scenario.
- Due to the complexity of SV comparison, evaluating methods against a single "ground truth" is insufficient.